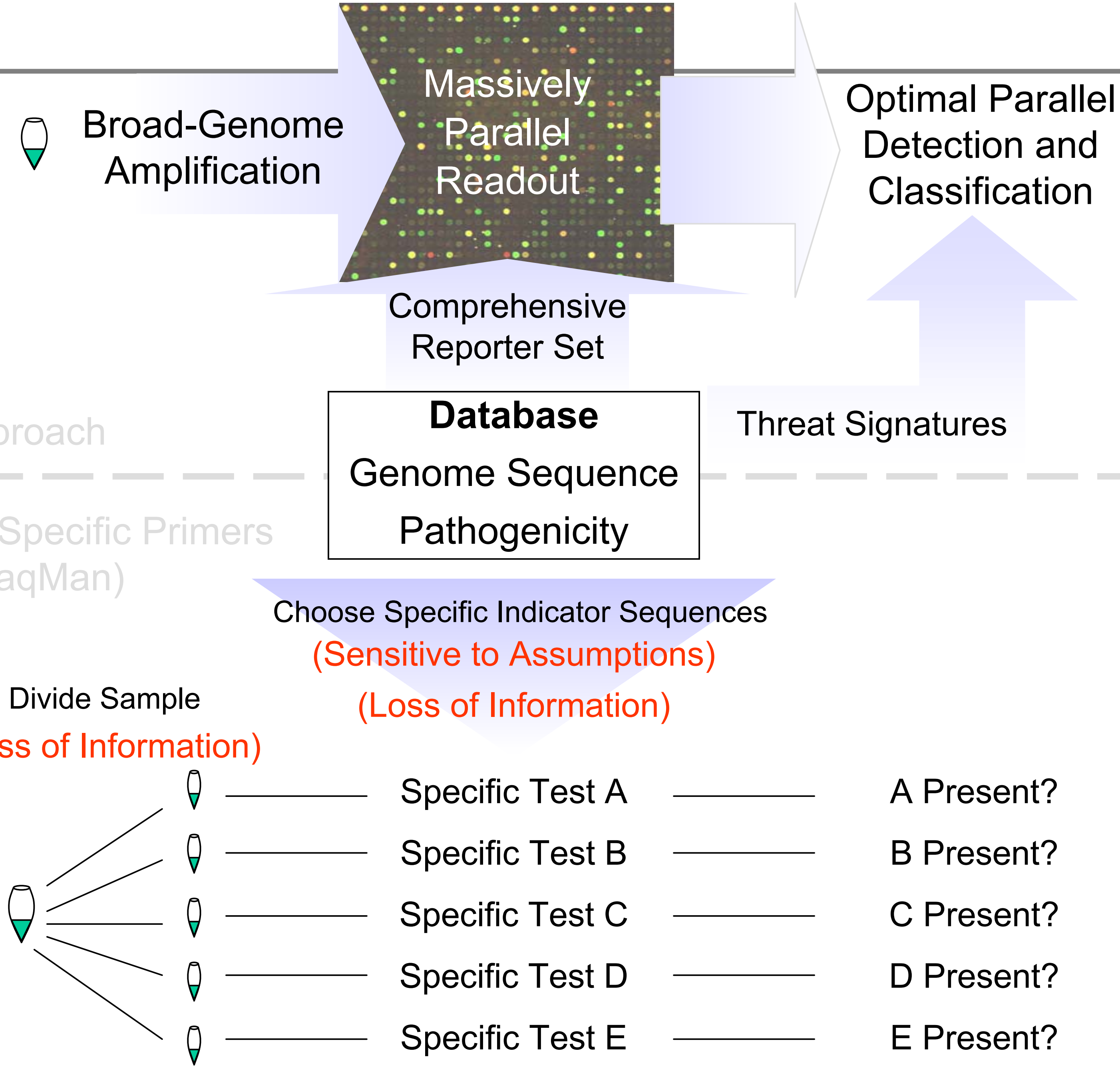
DNA Microarrays for Bioagent Detection

3 Feb 2004

Roland Stoughton – Genomic HealthCare

Cliff Lewis – SAIC

Mark Eshoo – Ibis Therapeutics



Broad-Genome Amplification

Massively Parallel Readout

Optimal Parallel Detection and Classification

Comprehensive Reporter Set

Database

Genome Sequence

Pathogenicity

Threat Signatures

Our Approach

Threat-Specific Primers (e.g., TaqMan)

Choose Specific Indicator Sequences

(Sensitive to Assumptions)

(Loss of Information)

Divide Sample

(Loss of Information)

Specific Test A — A Present?

Specific Test B — B Present?

Specific Test C — C Present?

Specific Test D — D Present?

Specific Test E — E Present?

DNA Microarrays for Bioagent Detection

- Microarray Probe Design Strategies
- Hybridization Model for Detection and System Simulation
- Detection Results for Bacteria
  - Detection of Individual Threats in Air Clutter Background
  - Parsing of the Anthracis Clade
- Summary

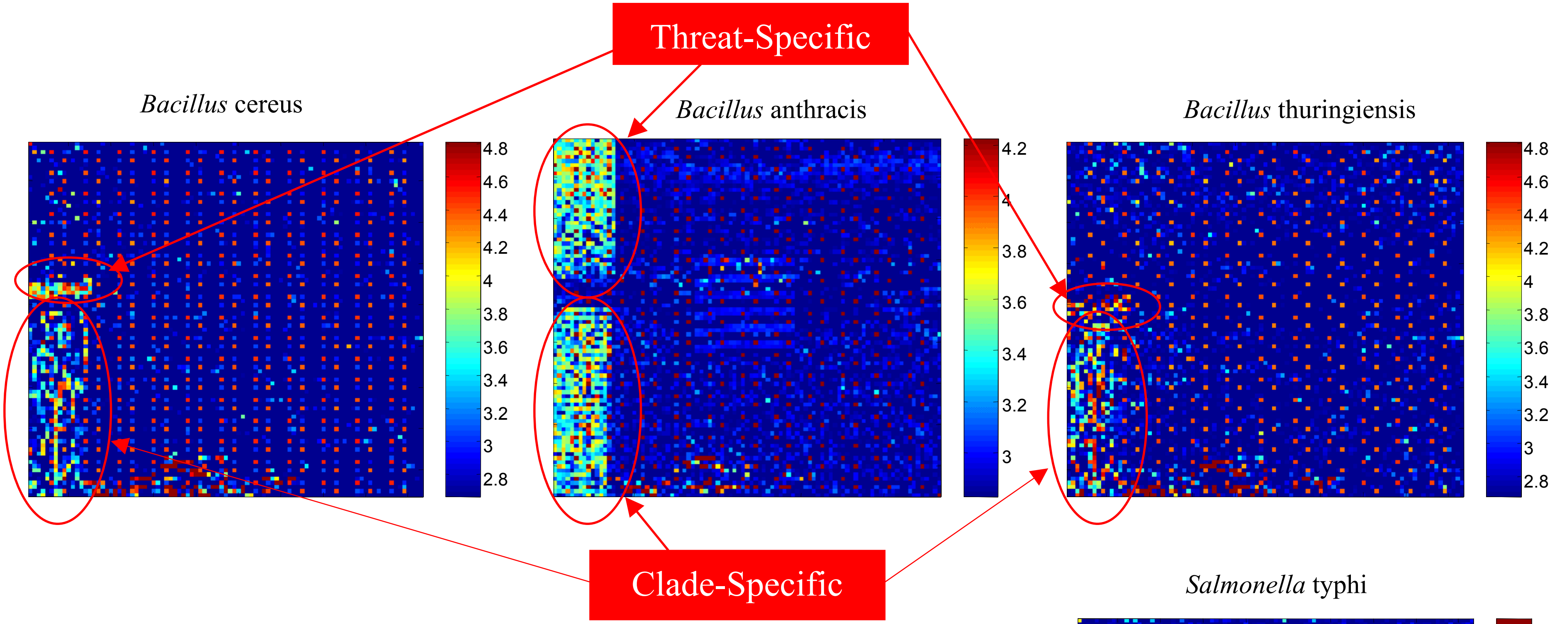
Probe Design Approaches

- How to detect all threats *and* distinguish near neighbors?

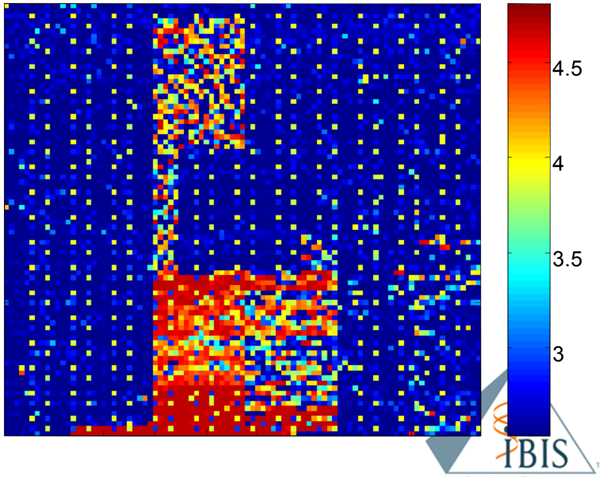
Conserved sequence regions are robust detectors, but don’t discriminate close neighbors.

Threat-unique sequences are good at discrimination, but may fail to detect due to strain variation or bioengineering.

Threat-Specific Probes for Resolution
Clade-Specific Probes for Robustness



Salmonella typhi



- **Agilent 30-mer oligo array, 8000 features**
- **6 threats, 8 near neighbors, in 6 clades**
- **Pure genomic DNA with whole-genome amplification**
- **Probe locations re-arranged for display**

Microarray Probe Design Strategies

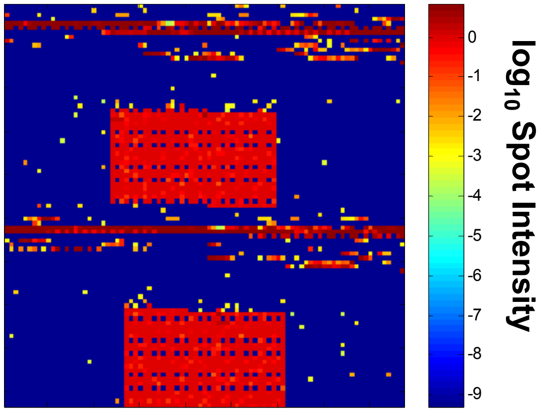
Name of Strategy	Amplification	Probe Design	Rationale	Probes included in Array Design	Detection Results Today
16S	PCR of 16S	16S Tiling	• Highly conserved • High sequence availability	X	
Clades	Whole Genome Amplification	Conserved regions for each Clade	• Unknown variants will be detected	X	X
Specific	Whole Genome Amplification	Organism-specific	• Ultimately best resolution • Virulence genes	X	X
Triangulation	PCR of Conserved Regions	Optimize primers and probes simultaneously	• Potentially the most efficient balance of primers and probes	X	

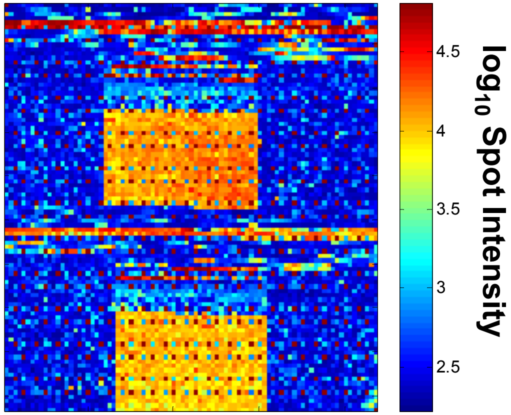
Detection Relies on Model for Threat Genome Hybridization

- Need to relate threat sequence and abundance to hyb intensity

- Model is crude but adequate
  - Assumes equilibrium
  - Different molecular species do not interfere
  - Based on ‘Nearest Neighbor’ quartet energies
  - Tuned to surface-phase hybridization

E Coli K-12 Signal Model





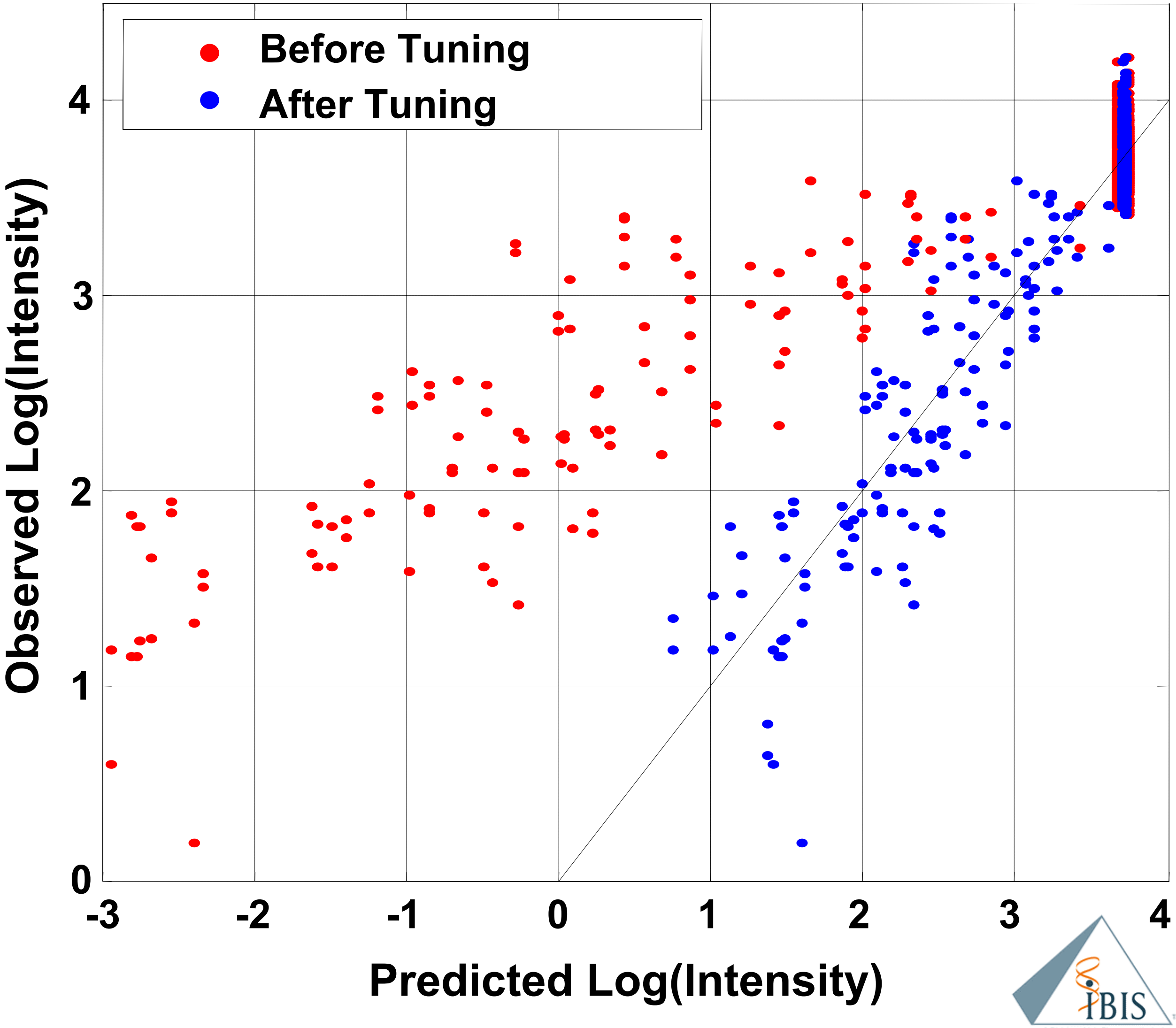
E Coli K-12 Microarray Hybridization

Nearest-Neighbor Model is Tuned for Surface-Phase Hybridization

- Model is adjusted to match observed hybridization of matched and mismatched duplexes

- Parameters fit
  - Energy penalty for sequence mismatch
  - Binding site density

Labeled 50-mer oligo spike-ins



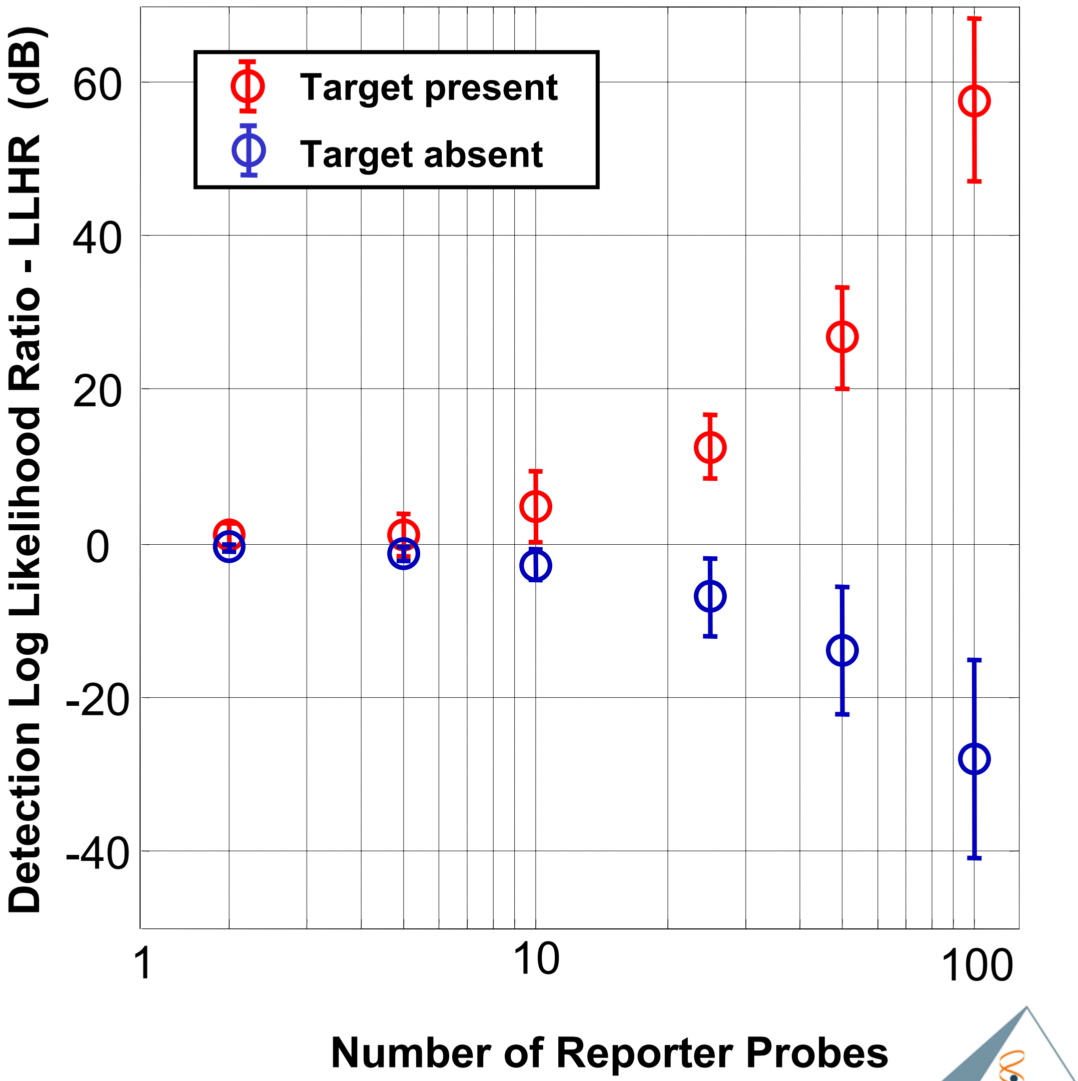
Having Many Probes per Threat Allows Robust Detection

Relies on model for threat genome hybridization

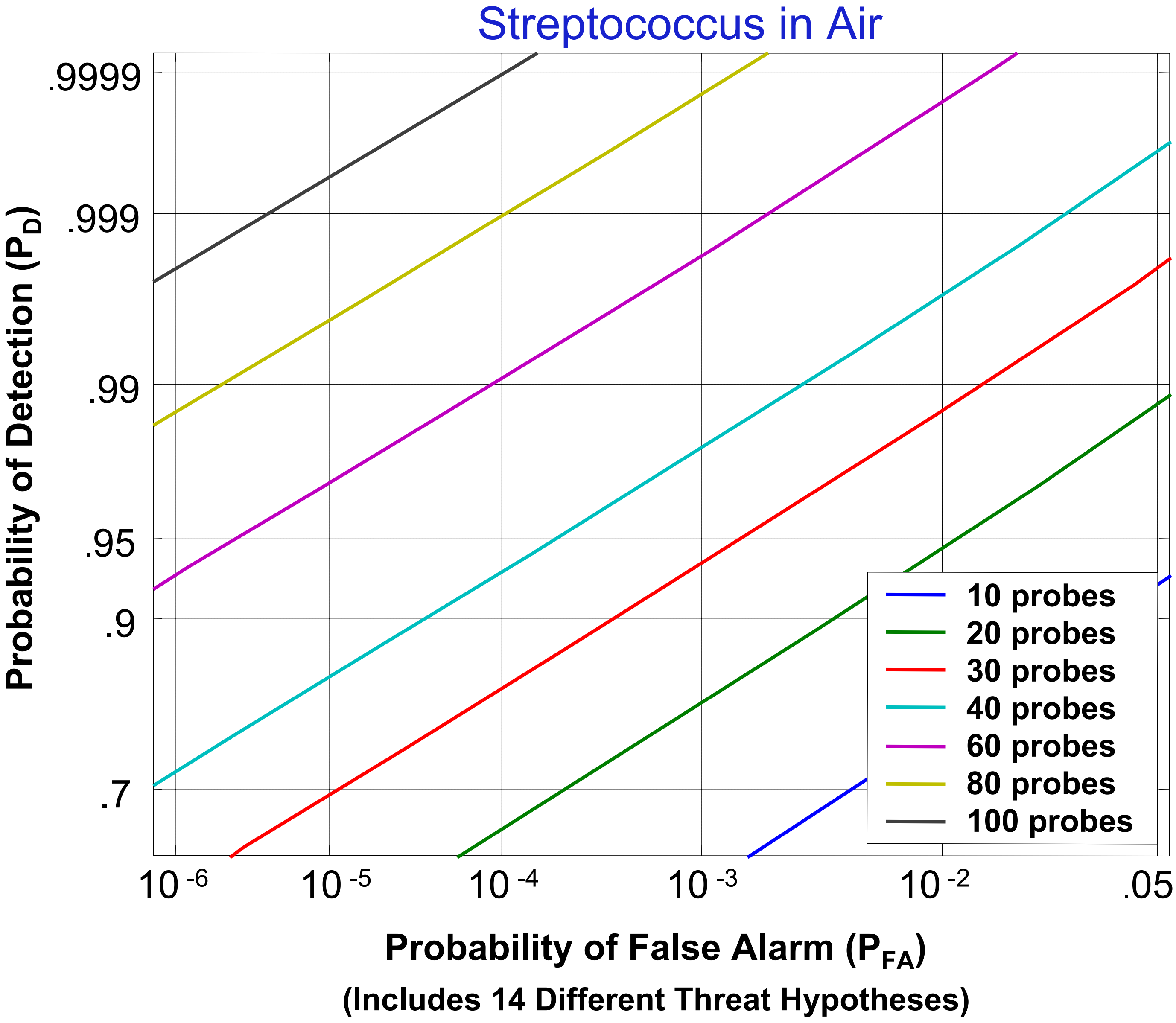
Likelihood Ratio based on number of probes brighter than 3x background

Low number of genome copies spiked into sample from 18,000 L of air

Streptococcus in Air



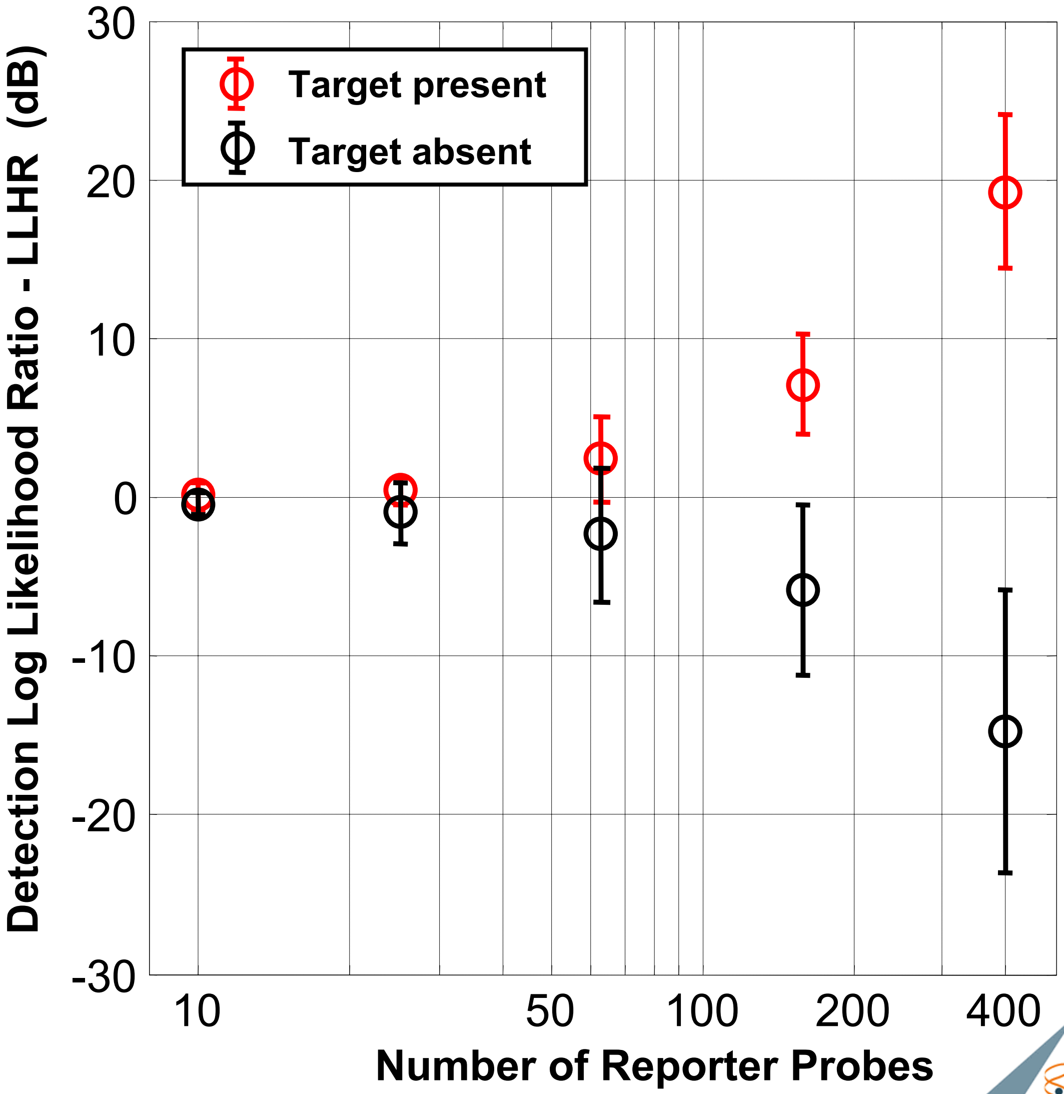
P_D - P_{FA} Performance Should Be Adequate for Monitoring Applications



A Division of Isis Pharmaceuticals, Inc.

Detection of Anthracis in Air

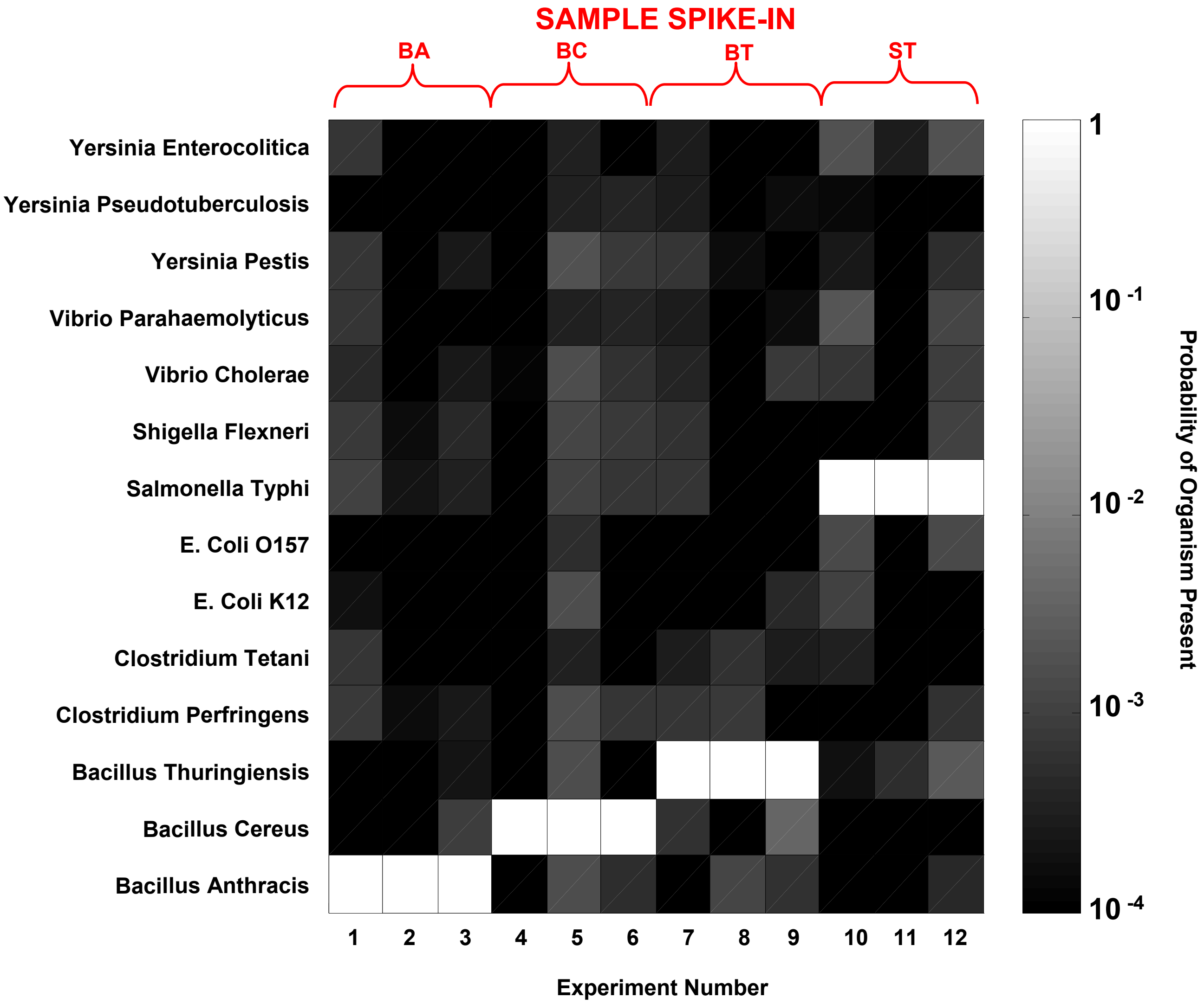
Low number of genome copies spiked into sample from 18,000 L of air



Probe Diversity Also Enables Discrimination Between Phylogenetic Near-Neighbors

All members of the anthracis clade are robustly and separately detected when analyzed in pure sample with whole genome amplification

Note: Probes were not required from PX01 or PX02 plasmids!



DNA Microarrays for Bioagent Detection

Summary

- Ability to dedicate many reporters to each threat provides robust detection and discrimination
- Clade-specific probes should provide additional robustness to strain variation and bioengineering
- Without much optimization, microarray sensitivity is already down to low number of genome copies
- DNA clutter from 18,000 L air sample (indoor collection) did not prevent robust detection of *Strep* and *B.* Anthracis

IBIS™
A Division of Isis Pharmaceuticals, Inc.

BACKUP

SAIC/Ibis Differs from LLNL Approach

• No sample division • Linear amplification	Potentially supports simultaneous quantitation of multiple threats down to a few genome copies, but amplification needs work
• Broad genome amplification	Will suffer more from background clutter
• Probes for conserved and for unique regions	Obtain best resolution and robustness to unknown variants

Derivation of $[L]_0$ (page 1/2)

(1) $L + R \overset{Ka}{\longleftrightarrow} D$ where *L* is ligand, *R* is receptor, *D* is the duplex, and K_a is the association constant at equilibrium

(2) $K_a = \dfrac{[D]}{[L][R]}$

(3) $R_0 = R + D = D(1 + \dfrac{R}{D})$ by mass conservation

(4) $\dfrac{[R]}{[D]} = \dfrac{1}{K_a[L]}$ from (2)

(5) $\dfrac{[R]}{[D]} = \dfrac{R}{D}$ since the volume of *R* is the same as *D*

(6) $R_0 = D(1 + \dfrac{1}{K_a[L]}) = D + \dfrac{D}{K_a[L]}$ combining (4) and (5) into (3)

(7) $[L] = \dfrac{D}{K_a(R_0 - D)}$ solving for $[L]$

K_a and R_0 are known, and *D* can be obtained through experimental result (8)

Derivation of $[L]_0$ (page 2/2)

(8) $$\frac{D}{R_0} = \frac{SI}{SI_{Max}} = f$$ where *f* is the fraction of receptor sites bound, and *SI* is signal intensity

(9) $$[L] = \frac{\frac{D}{R_0}}{K_a(1-\frac{D}{R_0})} = \frac{f}{K_a(1-f)} = \frac{\frac{SI}{SI_{Max}}}{K_a(1-\frac{SI}{SI_{Max}})} = \frac{SI}{K_a(SI_{Max}-SI)}$$ rearranging (7), and plugging in (8)

(10) $$V_L[L]_0 = V_L[L] + \frac{D}{N_{AV}}$$ by conservation of ligand mass, where V_L is the ligand (sample) volume and N_{AV} is Avogadro's number

(11) $$[L]_0 = \frac{SI}{K_a(SI_{Max}-SI)} + \frac{SI \times R_0}{SI_{Max} \times N_{AV} \times V_L}$$ combining (9) and (10)

A Division of Isis Pharmaceuticals, Inc.

LLHR Equations

Qualitative:

$$\Lambda(k) = \frac{\begin{array}{c}\text{If organism is present in sample, the probability of the SI of k out of n}\\ \text{probes being above SNR threshold } \tau \text{ given genomic representation r}\end{array}}{\begin{array}{c}\text{If organism is not present in sample, the probability of the SI of k out of n}\\ \text{probes being above SNR threshold } \tau \text{ given single probe false alarm rate } \mathrm{R}_{\mathrm{fa}}\end{array}}$$

Quantitative:

$$\Lambda(k) = \frac{p(H_1; r, n, k, \tau)}{p(H_0; R_{fa}, n, k, \tau)} = \frac{\binom{n}{k} r^k (1-r)^{n-k}}{\binom{n}{k} (R_{fa})^k (1-R_{fa})^{n-k}}$$

$$\log(\Lambda(k)) = k \log\left(\frac{r}{R_{fa}}\right) + (n-k)\log\left(\frac{1-r}{1-R_{fa}}\right)$$

$n =$ Number of organism reporter probes

$k =$ Number of probes with SNR above threshold τ

$r =$ Representation of genome in amplified sample

$R_{fa} =$ Probability of reporter probe exceeding SNR threshold due to clutter/noise

$\tau =$ Single probe SNR threshold

Detection Power and the F-Statistic

Yashar Behzadi
Nate Kowahl
Clifford Lewis
11/21/2003

The detection problem is stated as a choice between two hypotheses, defined in the terms of a general linear model:

1) H_0, y=Sb+n
   a. Null hypotheses where the signal of interest is not present
2) H_1, y=Xh+Sb+n
   a. Signal of interest is present

Where y is Nx1 vector of the observed data, X is an N x k design matrix, h is a k x 1 parameter vector, S is a N x 1 matrix consisting of nuisance model functions, b is a *l* x 1 vector of nuisance parameters, and n is a N x 1 vector that represents additive Gaussian noise.

A decision between these two hypotheses is made using the following definition of the general F-statistic.

$$F = \frac{N-k-l}{k} \frac{y^T P_{P_S^{\perp} X} y}{y^T (I - P_{XS}) y}$$

where P_{XS} is the projection onto the subspace <XS> and $P_{P_S^{\perp} X} = P_s^{\perp} X (X^T P_s^{\perp} X)^{-1} X^T P_s^{\perp}$ is the projection onto the signal subspace <X> that is orthogonal to the interference subspace <S>. The F-statistic is the ratio between the estimate of the average energy that lies in the part of the signal subspace <X> that is orthogonal to <S> and the noise variance derived from the energy in the data space that is not accounted for by the energy in the combined signal and interference subspace <XS>. For detection of a specific organism, the interference space is defined by a background DC term and by the modeled response of the microarray to the other possibly present organisms.

When the null hypothesis H_o is true, the F-statistic is governed by an F-distribution with k and N-k-*l* degrees of freedom. To use the F-statistic, we define a threshold α. If F > α, we choose H_1 otherwise we choose H_o. In order to define the value of α, we first choose a value for pFa. α is then calculated from the relationship 1-pFa=cdf(α). The F-statistic as applied to microarrays is summarized below in figure 1 and 2.